Problem A.: (50 points)

1. Let X and Y be independent random variables. What is the mean and variance of $X - Y$?

2. If X and Y are independent and $P(X) = 0.5$, what is of $P(X \mid Y)$?

3. If a researcher has set the significance level at 5% and the test statistic yields a p-value of 0.06, should the researcher should reject the null hypothesis.

4. If X has a normal distribution with mean 3 and standard deviation 5, what is the distribution of

$Z = \frac{X-3}{5}$?

5. If I want to double my certainty (make the confidence interval half the size), by how much I need to increase the sample size?

6. Manchester City (the EPL champions in 2017) are playing Arsenal. Suppose number of goals of two teams are independent. Both teams are expected on average to score two goals. What is the probability of a $1-1$ draw?. (Hint: Poisson model, density of Poisson(λ) distribution is $P(X=k) = e^{-\lambda}\lambda^k/k!$ where $k! = 1 \times 2 \times .. \times k$.)

7. A semi-conductor company knows from experience that 0.2% of chips will have imperfections. Suppose it makes 1000 such chips, what is the probability that **at least one** is imperfect?

8. Suppose that the annual returns for Tesla stock has a mean of 20% and a standard deviation of 10%. What distribution (normal/poisson/binomial) best describes the distribution of the returns? What is the probability that Tesla has returns greater than 20% for next year as predicted by your model?

9. A friend claims she can tell the difference between Evian and Dasani bottled water. Suppose p is the probability she can identify Evian correctly. In a random experiment with 100 repeated tests, the proportion that she can correctly identified the Evian water is $p = 0.6$. Formulate the null hypothesis and perform the hypothesis test at the 95% level.

Problem B: Bayes (20 points)

Shipments from an online retailer take between 1 and 7 days to arrive, depending on where they ship from, when they were ordered, the size of the item, etc.

Suppose the distribution of delivery times has the following distribution function:

x	1	2	3	4	5	6	7
P($X = x$)							
P($X \leq x$)	0.10	0.20	0.70	0.75	0.80	0.90	1

1. Fill in the above probability table.
2. What is the conditional probability of a delivery arriving on day four given that it did not arrive in the first three days? (Hint: find $P(X = 4 \mid X >= 4)$)

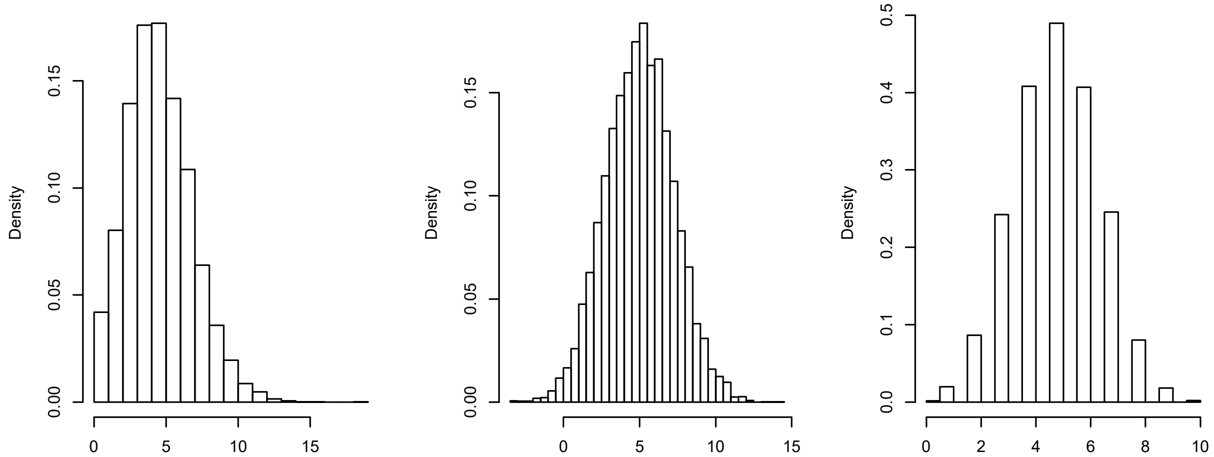
Problem C: A/B Testing. (20 points)

During a recent breakout of the flu, 850 out 6,224 people diagnosed with the virus presented severe symptoms.

During the same flu season, a experimental anti-virus drug was being tested. The drug was given to 238 people with the flu and only 6 of them developed severe symptoms.

Based on this information, can you conclude, for sure, that the drug is a success?

Problem D: Match the Distribution (20 points)



- Left histogram shows the data collected by the new coronavirus hospital in Wuhan. It shows the number of patients admitted every hour.
  a) What distribution would you use to describe this data?
  b) What is your best guess (based on the histogram) about parameter(s) of this distribution?

- Central histogram plots observed differences between the body temperature of the patients infected with coronavirus from the body temperature of a healthy person (98.6 F).
  c) What distribution would you use to describe this data?
  d) What is your best guess (based on the histogram) about parameter(s) of this distribution?

- You split checked-in patients into groups of 10. The right histogram shows distribution of the patients admitted to the hospital form each group of 10. The rest are not actually sick and get sent home.
  e) What distribution would you use to describe this data?
  f) What fraction of the patients get admitted?
  g) What is the probability that two patients out of 10 get admitted?

Problem E: Russian Parliament Election Fraud. (15 points) On September 28, 2016 United Russia party won a supermajority of seats, which will allow them to change the Constitution without any votes of other parties. Throughout the day there were reports of voting fraud including video purporting to show officials stuffing ballot boxes. Additionally, results in many regions demonstrate that United Russia on many poll stations got anomalously closed results, for example, 62.2% in more than hundred poll stations in Saratov Region.

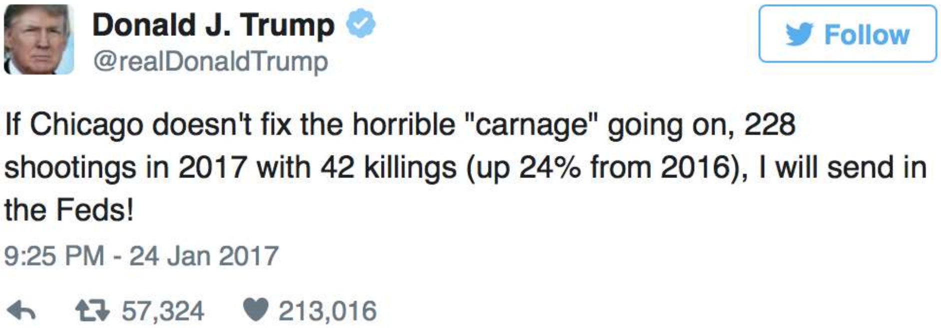
Using assumption that United Russia's range in Saratov was [57.5%, 67.5%] and results for each poll station are rounded to one decimal point (when measure in percent), calculate probability that in 100 poll stations out of 1800 in Saratov Region the majority party got exactly 62.2%.

Do you think it can happen by a chance?

Hint: Assume that prob of any 1 decimal place number b/w 57.5 and 67.5 is equal.

Problem F: Chicago Crime Data Analysis. (30 points)

On January 24, 2017 Donald Tramp tweeted about ”horrible” murder rate in Chicago.



Donald J. Trump
@realDonaldTrump

If Chicago doesn't fix the horrible "carnage" going on, 228 shootings in 2017 with 42 killings (up 24% from 2016), I will send in the Feds!

9:25 PM - 24 Jan 2017

57,324 213,016

Our goal is to analyze the data and check how statistically significant such a statement. I downloaded Chicago’s crime data from the data portal: `data.cityofchicago.org`. This data contains reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department’s CLEAR (Citizen Law Enforcement Analysis and Reporting) system. In order to protect the privacy of crime victims, addresses are shown at the block level only and specific locations are not identified. This data set has 6.3 million records. Each crime incident is categorized using one of the 35 primary crime types: NARCOTICS, THEFT, CRIMINAL TRESPASS, etc.. I frittered incidents of type HOMICIDE into a separate data set stored in `chi_homicide.rds`. Use `chi_crime.R` as a starging script for this problem.

a) Look at the of the homicide incidents. You will see similar picture as you saw with the heat plot. There is an ”island” with no homicide incidents the south side Chicago! Can you explain why?

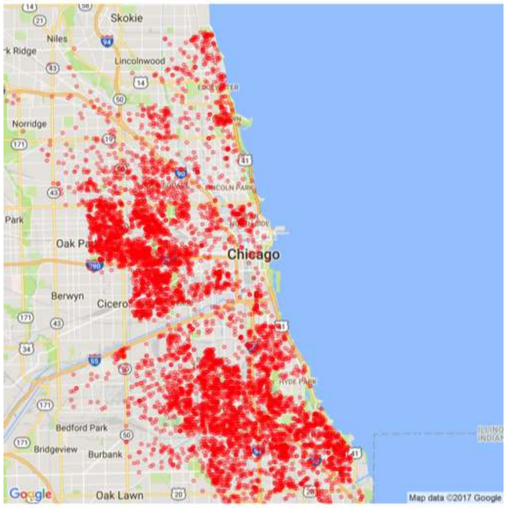


Figure 1: Homicide Map

b) Though president's tweet is consistent with the data (`goo.gl/VTPzFw`), observing 52 homicides in January is not that unusual. Calculate the total number of homicides for each January. Estimate 95% confidence interval for the mean μ over January homicides. Is 52 within the interval?

c) The history of 2001-present data is rather short. Chicago tribune provided total number of homicides for Chicago for each month of the 1957-2014 period. Use this data set and calculate the confidence interval for μ. Further answer the following questions: (i) Assuming monthly homicide rate follows Normal distribution, what is the probability that we observe 52 homicides or more? (ii) Do you think Normality assumption is valid? (iii) Assuming monthly homicide rate follows Poisson distribution, what is the probability that we observe 52 homicides or more?

d) There is another hypothesis that rise in murder is related to the pullback in proactive policing that started in November of 2015 as a result of Laquan McDonald video release (`https://goo.gl/7cm1CC`, `https://goo.gl/WcH2uB`). I calculated total number of homicides for each day and split data into two parts: before and after video release. Using t-ratio, check the hypothesis H_0: the homicide rate did not change after video release.

Problem G: A/B Testing for Search Algorithm. (30 points)

Use dataset from `ab_browser_test.csv`

Here is the definition of the columns:

- `userID`: unique user ID
- `browser`: browser which was used by userID
- `slot`: status of the user (exp = saw modified page, control = saw unmodified page)
- `n_clicks`: number of total clicks user did during as a result of `n_queries`
- `n_queries`: number of queries made by `userID`, who used browser `browser`
- `n_nonclk_queries`: number of queries that did not result in any clicks

Note, that not everyone uses a single browser, so there might be multiple rows with the same `userID`. In this dataset combination of `userID` and `browser` is the unique row identifier.

a) Count how many users in each group. How much larger (in percent) exp group when compared to control group

b) Construct 95% confidence interval for means of number of clicks in group exp and group control. Are the means are significantly different?

c) Use z-ratio for the means, to perform hypothesis testing, with H_0: there is no difference in average number of clicks between 2 groups